

# Moving Taylor Bayesian Regression for nonparametric multidimensional function estimation with possibly correlated errors

Jobst Heitzig

Potsdam Institute for Climate Impact Research (PIK), Transdisciplinary Concepts and Methods,  
P.O. Box 60 12 03, 14412 Potsdam, Germany.

**Summary.** We present a nonparametric method for estimating the value and several derivatives of an unknown, sufficiently smooth real-valued function of real-valued arguments from a finite sample of points, where both the function arguments and the corresponding values are known only up to measurement errors having some assumed distribution and correlation structure. The method, *Moving Taylor Bayesian Regression* (MOTABAR), uses Bayesian updating to find the posterior mean of the coefficients of a Taylor polynomial of the function at a moving position of interest. When measurement errors are neglected, MOTABAR becomes a multivariate interpolation method. It contains several well-known regression and interpolation methods as special or limit cases. We demonstrate the performance of MOTABAR using the reconstruction of the Lorenz attractor from noisy observations as an example.

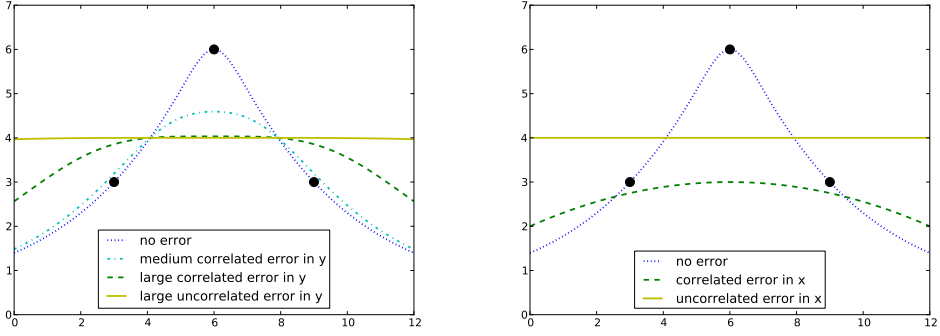
**Keywords:** interpolation; irregular sampling; numerical differentiation; rational function; smoothing

## 1. Introduction

### 1.1. Motivation

The basic task of regression analysis – the estimation of values of an unknown, sufficiently smooth function  $f$  of one or several real arguments, given a finite amount of possibly noisy data – occurs pervasively in many kinds of quantitative scientific research. Most existing approaches to this task can roughly be classified into two groups: *parametric* or model-based approaches such as polynomial regression, spline smoothing (Reinsch, 1967), or kriging (Krige, 1952), and *nonparametric* approaches such as local regression, nearest or natural neighbour estimation (Sibson, 1981), inverse distance weighting (Shepard, 1968), or kernel smoothing. Although parametric methods are usually based on more rigorous reasoning such as maximum likelihood estimation, Bayesian updating, approximation theory, or some other kind of optimization, their results are only guaranteed to be reliable if the sought function is assumed to belong to some particular model or function class with a small number of parameters, e. g., polynomials or splines of a fixed degree and fixed set or number of break points. If, as is often the case in empirical research, this strong assumption can not be justified, nonparametric methods are available which, however, are usually based on various forms of more heuristic reasoning, frequently involve the choice of several control parameters such as the number of neighbours or the choice of a weight or kernel function, and sometimes do not provide an easily interpreted assessment of reliability such as a standard error or confidence interval. Also, many regression methods are not or only restrictedly applicable if one or several of the following conditions apply: (i) there might be measurement errors not only in the function values (the “dependent” variable) but also in the function arguments (the “independent” variables), (ii) measurement

E-mail: heitzig@pik-potsdam.de



**Fig. 1.** (Colour online) Illustration of the effect of error correlation on some MOTABAR estimates.

errors might not be independent and identically distributed, (iii) the function is of more than one real argument, (iv) the measured sample is irregularly distributed in the function's argument space, and/or (v) also several derivatives of the function shall be estimated.

In order to illustrate how plausible estimates can depend on the assumed amount and correlation of measurement errors, consider the minimal data  $(3, 3)$ ,  $(6, 6)$ , and  $(9, 3)$  and assume that we want to estimate  $f$  and  $f'$  on the interval  $[0, 12]$  on the basis of this data. In the left diagram in Fig. 1, four different such estimates of  $f$  are shown, which were produced with the method we will describe in this article, using different assumptions on the measurement error contained in the data. Without measurement errors, i. e., if both  $x$  and  $y$  data are precise, it is plausible to estimate  $f(3) = 3$  and  $f'(3) \approx 1$ , as on the blue dotted line. Without errors in  $x$  but with large independent errors in  $y$ , the apparent slope becomes quite uncertain, and one would rather estimate  $f(3)$  by the sample mean, 4, as on the yellow line. However, if the errors in  $y$  are highly correlated, e. g., because of a systematic but unknown bias in the measurement equipment, then one knows that the data is basically only shifted in the  $y$  direction, which has no influence on slopes, so one would probably expect  $f'(3) > 0$  again, as on the green dashed line. The smaller the errors, the more the slope should resemble 1, as on the cyan dash-dotted line. Similarly, if the  $y$  data is precise but the  $x$  data is highly uncertain, one would use the sample estimate if errors are uncorrelated, as on the yellow line in the right diagram in Fig. 1. If errors in  $x$  are correlated, one would retain some slope information which suggests that the values to the left and right of the three data points are rather below than above three. Hence the estimate would be lower than with uncorrelated errors, as on the green dashed line. The occurrence of this lowering effect also shows that the effect of errors in the  $x$  dimension is usually not the same as the effect of some “equivalent” amount of errors in the  $y$  dimension.

### 1.2. Moving Taylor Bayesian Regression

In this article, we develop a regression method which is nonparametric in the sense that it can estimate any sufficiently smooth function, but is still based on rigorous statistical reasoning, can deal with any of the above situations (i)–(v), and contains several existing methods as special or limiting cases. The method, *Moving Taylor Bayesian Regression (MOTABAR)*, is based on the following ideas:

- (a) Approximate  $f$  locally by a Taylor polynomial at a moving position of interest  $\xi$ .

- (b) Treat the unknown Taylor coefficients as parameters in a statistical model.
- (c) Use the measured data to update prior beliefs about these parameters using Bayesian updating.
- (d) Use the posterior mean (or mode) and variance of the parameters as estimates of the function's value and derivatives and the corresponding estimation uncertainty.

In principle, these steps can be performed for any form of error and prior distribution, but the method becomes particularly transparent if the value measurement error is assumed to be multivariate Gaussian and the involved prior distributions are also multivariate Gaussian. In that case, the resulting posterior distributions can be written as mixtures of Gaussians whose mean and variance can be derived analytically, and the resulting MOTABAR estimator  $\hat{f}(\xi)$  of  $f(\xi)$  turns out to be a mixture of smooth rational functions of  $\xi$  that can be computed using simple linear algebra. The mixing is related to the distribution of argument measurement errors. Without argument measurement errors,  $\hat{f}(\xi)$  is a smooth rational function of  $\xi$  whose algebraic form can be seen as a generalization of ordinary least squares regression estimators. For these reasons, we will focus on the case of Gaussian priors and value errors in this article.

The input data needed to apply MOTABAR are then

- the measured data,
- variances and possibly covariances of both argument and value measurement errors,
- a prior covariance matrix for the value of  $f$  and its derivatives of order  $< p$  for some  $p > 0$ ,
- prior variances for  $f$ 's derivatives of order  $p$ ,
- and, as the only free control parameter, an integer  $p$  larger than the order of any derivative of  $f$  one wants to estimate.

For particular cases of error and prior distributions, it will turn out that the MOTABAR estimate equal or approximate the results of some well-known other methods, including ordinary polynomial regression, inverse distance weighting, and linear interpolation on regular grids. This behaviour of our estimator is similar to that of a related approach by Wang et al. (2010a) in which  $f$  is also approximated by a Taylor polynomial at a moving position of interest  $\xi$ , but in which the Taylor coefficients are then estimated not by Bayesian updating but by minimizing a heuristic loss function motivated by approximation theory (see Sec. 5). Other cases of error and prior distributions result in plausible generalizations or variants of well-known methods, e.g., a new form of local polynomial interpolation and a Bayesian variant of inverse distance weighted smoothing.

In practise, the needed prior and error variances and covariances might themselves be estimated from other or even the same data, a question we do however not address in detail in this article.

*Comparison with other methods.* MOTABAR can be interpreted as a kind of *local* regression since although it takes into account also data points far away from  $\xi$ , it gives them much less influence on the estimate than those close to  $\xi$ . This fact is reflected in the occurrence of a *weight matrix*  $W$  in the estimator equation. But unlike other local or piece-wise methods such as nearest or natural neighbour regression, locally weighted scatterplot smoothing (LOESS) (Cleveland, 1979), or splines, the MOTABAR estimate  $\hat{f}(\xi)$  is infinitely smooth (i. e., infinitely often differentiable), at least as long as the value measurement errors have nonzero variance, and there are either no argument measurement errors or their probability density is infinitely smooth (e. g., when argument measurement errors are Gaussian as well). This is because in MOTABAR the weight of each individual data point in the

estimate depends smoothly on its distance from  $\xi$ , whereas in other methods the weights can switch from zero to nonzero in a nonsmooth way as  $\xi$  moves.

The nonsmooth change in weights that other methods involve is also counter-intuitive when arguments can only be measured with some error. E. g., suppose that measurements resulted in the argument-value pairs  $(-1, 0)$ ,  $(0, 1)$ ,  $(1, 0)$ , and  $(1/100, -1)$ , where the argument measurements involved an error of magnitude  $1/10$ , so that the last measurement might actually reflect  $f(-1/100)$  instead of  $f(1/100)$ . A linear interpolant of the four measurements would have a slope of  $\approx +1$ , but when the  $(1/100, -1)$  measurement is moved towards  $(1/100, -1)$ , the slope would discontinuously switch to  $\approx -1$ . A similar effect occurs with splines. In other words, when the ranking of the argument measurements with respect to their distance from  $\xi$  is uncertain due to measurement errors, it seems inappropriate to use a method that strongly relies on the correctness of this ranking, e. g., by using only the  $k$  nearest neighbours of  $\xi$ ; although disregarding faraway measurements or extreme observations completely instead of just downweighting them might still be advisable to increase the robustness of the method if outliers might exist, e. g., due to fat-tailed error distributions.

One existing class of methods, Inverse Distance Weighting (IDW) (Shepard, 1968), also uses smoothly decaying weights that only depend on the distance from  $\xi$ . In a commonly used variant of IDW, the weights are inversely proportional to the square or a larger power of the distance, and we will show that this method can be derived as a special case of MOTABAR in which a noninformative prior distribution for  $\varphi$  is used.

A common feature of many interpolation and smoothing methods, including IDW and other weightings- or kernel-based methods, linear interpolation, and nearest or natural neighbour interpolation, is that the estimate cannot exceed the largest measured values, even if the data strongly suggest a nonzero slope at the largest data point. Such methods will therefore always underestimate the maxima of  $f$ . Other methods, like polynomial regression, can in some situations ‘overshoot’ and result in estimates that lie far outside the measured range. Spline methods are often considered a good compromise between these two behaviours regarding maxima. Depending on the choice of  $p$ , MOTABAR will behave quite similar to spline methods in this respect. As a consequence, the weights of individual data values in the estimate might be  $< 0$  or  $> 1$ .

Similar to other nonparametric methods or methods with many parameters, the MOTABAR estimate might fit the data too narrowly and show too much fluctuations due to this ‘overfitting’ when  $p$  and the prior variances are badly chosen. If the prior variances for higher derivatives grow too fast, the estimate is allowed to vary on smaller scales than the sampling density can resolve reliably in view of the assumed error distributions. The amount of overfitting can thus be controlled by varying  $p$  and the priors.

### 1.3. Framework

Assume that we are interested in the value of a certain function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  and all its derivatives of order  $< p$  at a certain *position of interest*  $\xi \in \mathbb{R}^d$ , where  $p > 0$  and  $f$  is assumed to be  $p$  times continuously differentiable. We take the convention to write elements of  $\mathbb{R}^d$  as column vectors and to enumerate their  $d$  components with a subscript index, hence  $\xi = (\xi_1, \dots, \xi_d)'$  where the  $'$  symbol denotes transposition. Assume that the only information we have about  $f$  is (i)  $N$  *measured data points*  $(x^{(\hat{i})}, y^{(\hat{i})})$  with  $i = 1 \dots N$ , (ii) some estimates of the magnitude of *measurement errors*, and (iii) some beliefs about the variability of  $f$  and some of its derivatives. These assumptions will be made more precise later. We enumerate measurements with superscript indices in guillemots, hence  $x^{(\hat{i})} = (x_1^{(\hat{i})}, \dots, x_d^{(\hat{i})})'$ . Assume the  $i$ -th data point  $(x^{(\hat{i})}, y^{(\hat{i})})$  is the result of trying to measure  $f$  at the argument  $x^{(\hat{i})}$ , but the measurement might involve an error in both the argument and the value, so that the actual result  $y^{(\hat{i})}$  is the value of  $f$  at a slightly different argument  $\chi^{(\hat{i})} = x^{(\hat{i})} - \gamma^{(\hat{i})}$  (where

$\gamma^{i\flat} \in \mathbb{R}^d$  is the *argument error*) plus some *value error*  $\varepsilon^{i\flat} \in \mathbb{R}$ . In other words,

$$y^{i\flat} = f(x^{i\flat} - \gamma^{i\flat}) + \varepsilon^{i\flat}. \quad (1)$$

**Multi-index notation.** For dealing with higher-order derivatives and multidimensional Taylor polynomials of  $f$ , it is convenient to use a *multi-index*  $\alpha \in \mathbb{N}^d$  consisting of nonnegative integers  $\alpha_k \geq 0$  for  $k = 1 \dots d$ . We then use the denotations

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)', \quad |\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_d \in \mathbb{N}, \quad (2)$$

$$\alpha! = \alpha_1! \alpha_2! \dots \alpha_d! \in \mathbb{N}, \quad x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \in \mathbb{R}, \quad (3)$$

$$(D^\alpha f)(x) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \dots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f(x_1, \dots, x_d) \in \mathbb{R}, \quad (4)$$

where  $x \in \mathbb{R}^d$  and  $|\alpha| \leq p$ . Note that the order of differentiation in  $(D^\alpha f)(x)$  is unimportant for  $|\alpha| \leq p$ . Our quantities of interest are then the derivatives

$$\varphi^{(\alpha)} = (D^\alpha f)(\xi) \quad (|\alpha| < p). \quad (5)$$

## 2. Using Taylor's Theorem to get a local model of the function at the position of interest

For a given position of interest  $\xi \in \mathbb{R}^d$ , a relationship between our quantities of interest  $\varphi^{(\alpha)}$  and the data  $x^{i\flat}, y^{i\flat}$  is given by Taylor's Theorem, which leads to

$$y^{i\flat} = \sum_{\alpha, |\alpha| < p} X^{i, \alpha} \varphi^{(\alpha)} + r^{i\flat} + \varepsilon^{i\flat}, \quad (6)$$

where

$$X^{i, \alpha} = \frac{(x^{i\flat} - \xi)^\alpha}{\alpha!}, \quad r^{i\flat} = \sum_{\alpha, |\alpha| = p} X^{i, \alpha} \psi^{(\alpha i)}, \quad (7)$$

$$\psi^{(\alpha i)} = (D^\alpha f)(\xi + \lambda^{(\alpha i)}(x^{i\flat} - \xi)), \quad \lambda^{(\alpha i)} \in [0, 1] \quad (|\alpha| = p). \quad (8)$$

Note that there are  $m = \binom{p-1+d}{d}$  many choices for  $\alpha$  with  $|\alpha| < p$  and  $\binom{p-1+d}{d-1}$  many choices for  $\alpha$  with  $|\alpha| = p$ . It will be convenient to stack all relevant quantities into the column vectors

$$\mathbf{x} = (x_1^{(1)}, \dots, x_d^{(1)}, \dots, x_1^{(N)}, \dots, x_d^{(N)})', \quad (9)$$

$$\boldsymbol{\gamma} = (\gamma_1^{(1)}, \dots, \gamma_d^{(1)}, \dots, \gamma_1^{(N)}, \dots, \gamma_d^{(N)})', \quad (10)$$

$$\mathbf{X} = \mathbf{x} - \boldsymbol{\gamma}, \mathbf{y} = (y^{(1)}, \dots, y^{(N)})', \boldsymbol{\varepsilon} = (\varepsilon^{(1)}, \dots, \varepsilon^{(N)})', \mathbf{r} = (r^{(1)}, \dots, r^{(N)})',$$

$$\boldsymbol{\varphi} = (\varphi^{(\alpha)} : |\alpha| < p)', \quad \boldsymbol{\psi} = (\psi^{(\alpha i)} : |\alpha| = p, i = 1 \dots N)', \quad (11)$$

and the matrix

$$\mathbf{X} = (X^{i, \alpha} : i = 1 \dots N, |\alpha| < p). \quad (12)$$

Note that the column vector  $\boldsymbol{\psi}$  has a row for each combination of  $\alpha$  and  $i$ , and we write this row index as  $\alpha i$ . This must not be confused with the row-and-column index pair  $i, \alpha$  of the matrix  $\mathbf{X}$ .

In other words,  $\varphi$  is an  $m \times 1$  vector,  $\psi$  is  $\binom{p-1+d}{d-1}N \times 1$ , and  $X$  is an  $N \times m$  matrix. Eq. 6 is now summarised in matrix notation as

$$\mathbf{y} = X\varphi + (\mathbf{r} + \varepsilon). \quad (13)$$

Although this is formally a linear regression model, we are not interested in estimating its coefficient matrix  $X$  (which we know already up to some measurement errors), but in estimating the regressor  $\varphi$ , and this we need to do for each position of interest  $\xi$  separately. This is the reason why we next apply Bayes' Theorem to Eq. 13.

### 3. Using Bayesian updating to estimate the function value and derivatives

#### 3.1. General approach

Let us model our information about the quantities of interest  $\varphi^{(a)} = (D^a f)(\xi)$  and about the other unknown terms in Eq. 13 as Bayesian beliefs, i. e., in the form of (subjective) probability distributions, by treating  $\varphi$  and  $\psi$  as random variables with Lebesgue-integrable densities  $\varrho(\cdot)$ . Because in general, each  $\psi^{(a,b)}$  corresponds to an argument  $\xi + \lambda^{(a,b)}$  that is different from  $\xi$ , we assume  $\varphi$  and  $\psi$  are independent. Also, we assume that the function-related quantities  $\varphi$  are independent from the measurement-related quantities  $\mathbf{x}$  and  $\gamma$ . The measured data  $\mathbf{x}, \mathbf{y}$  then allow us to update our prior beliefs  $\varrho(\varphi)$  about  $\varphi$  by applying Bayes' Theorem to Eq. 13, which leads to posterior beliefs

$$\varrho(\varphi|\mathbf{x}, \mathbf{y}) \propto \varrho(\varphi) \int_{\mathbb{R}^{Nd}} d^{Nd} \gamma \varrho(\gamma|\mathbf{x}) \int_{\mathbb{R}^N} d^N \mathbf{r} \varrho(\mathbf{r}|\mathbf{x}, \gamma) \varrho(\mathbf{y}|\mathbf{x}, \gamma, \varphi, \mathbf{r}). \quad (14)$$

Note that we follow the general convention to use the same symbol, here  $\varrho$ , to refer to all occurring probability densities since it will always be clear from the context which variable's density is meant. Also, we suppress the domain of integration in the following, and we are not interested in multiplicative constants that do not depend on  $\varphi$ , and use the symbol  $\propto$  to denote equality up to such a constant. To utilise Eq. 14, we need to specify

- a distribution  $\varrho(\gamma|\mathbf{x})$  for the argument errors that might depend on the arguments,
- a distribution  $\varrho(\varepsilon|\mathbf{x}, \gamma)$  of the value errors that might depend on the arguments and argument errors, and
- prior distributions  $\varrho(\varphi)$  and  $\varrho(\psi)$  expressing our initial information about the values and derivatives of  $f$  at  $\xi$  before the measurements.

The term  $\varrho(\mathbf{r}|\mathbf{x}, \gamma)$  can then be determined from  $\varrho(\psi)$  using Eq. 7, while the term  $\varrho(\mathbf{y}|\mathbf{x}, \gamma, \varphi, \mathbf{r})$  can be determined from  $\varrho(\varepsilon|\mathbf{x}, \gamma)$  using Eq. 13. As an estimate of  $\varphi$  one can then use any measure of central tendency of the posterior distribution as given by Eq. 14, e. g., the posterior mean, median, or mode, while an estimate of the estimation error would be given by a suitable measure of dispersion, e. g., the posterior variance or the median distance from the median. Although this strategy can in principle be applied to error and prior distributions of any form, our approach becomes especially simple in the Gaussian case.

### 3.2. Gaussian value error and priors

For the rest of this article, we will assume that value errors and priors are (multivariate) Gaussian,

$$\varrho(\varepsilon|\mathbf{x}, \boldsymbol{\gamma}) \propto \exp(-\varepsilon' P_\varepsilon \varepsilon/2), \quad (15)$$

$$\varrho(\boldsymbol{\varphi}) \propto \exp\{-(\boldsymbol{\varphi} - \boldsymbol{\mu}_\varphi)' P_\varphi (\boldsymbol{\varphi} - \boldsymbol{\mu}_\varphi)/2\}, \quad (16)$$

$$\varrho(\boldsymbol{\psi}) \propto \exp\{-(\boldsymbol{\psi} - \boldsymbol{\mu}_\psi)' P_\psi (\boldsymbol{\psi} - \boldsymbol{\mu}_\psi)/2\}, \quad (17)$$

where we use *precision matrices*  $P_\cdot = \Sigma_\cdot^{-1}$  and omit to denote the dependency on  $\mathbf{x}, \boldsymbol{\gamma}$  to simplify the following equations. Note that  $P_\varepsilon$  is a  $Nd \times Nd$  square matrix whose rows and columns we address using indices of the form  $ik$  with  $i \leq N$  and  $k \leq d$ . The case in which it is known that there are no value errors requires some special treatment since then  $\Sigma_\varepsilon = 0$  so that  $P_\varepsilon$  does not exist. In that case,  $\varrho(\varepsilon|\mathbf{x}, \boldsymbol{\gamma})$  behaves like a Dirac delta function, giving  $\int d^N \varepsilon \varrho(\varepsilon|\mathbf{x}, \boldsymbol{\gamma}) g(\varepsilon) = g(0)$  for all integrable functions  $g$  of  $\varepsilon$ .

The Taylor remainders  $\mathbf{r}$  given  $\mathbf{x}$  and  $\boldsymbol{\gamma}$  are then also Gaussian with mean  $\boldsymbol{\mu}_r$  and covariance matrix  $\Sigma_r$  given by

$$\boldsymbol{\mu}_r^{i\flat} = \sum_{\substack{\alpha \\ |\alpha|=p}} X^{i,\alpha} \boldsymbol{\mu}_\psi^{i,\alpha}, \quad \Sigma_r^{i,j\flat} = \sum_{\substack{\alpha \\ |\alpha|=p}} \sum_{\substack{\beta \\ |\beta|=p}} X^{i,\alpha} \Sigma_\psi^{\alpha i, \beta j} X^{j,\beta}. \quad (18)$$

Note that for almost all choices of  $\boldsymbol{\gamma}$  and  $\Sigma_\psi$ , the matrix  $\Sigma_r$  is nonsingular. We will therefore assume that  $\boldsymbol{\gamma}$  has a continuous distribution and  $\Sigma_\psi$  is nonsingular, so that almost surely  $\Sigma_r$  is nonsingular. If it is known that  $\boldsymbol{\gamma} = 0$ , we assume instead that  $\mathbf{x}, \xi$  are in general position, in which case  $\Sigma_r$  is also nonsingular.

To get an understanding of the relative sizes of the entries in  $\Sigma_r$ , consider the special case in which  $\Sigma_\psi$  is block-diagonal with identical  $N \times N$  blocks  $V$ , so that  $\Sigma_\psi^{\alpha i, \beta j} = \delta_{\alpha\beta} V^{i,j}$ . Then  $\Sigma_r^{i,j\flat} = V^{i,j} \{(\chi^{i\flat} - \xi)'(\chi^{j\flat} - \xi)\}^p / p! \neq 0$ , showing how the covariance of the remainders  $r^{i\flat}$  and  $r^{j\flat}$  grows with the  $p$ -th power of the scalar product of  $\chi^{i\flat}$  and  $\chi^{j\flat}$ . If  $V$  is diagonal, also  $\Sigma_r$  is diagonal and the variance of the remainder  $r^{i\flat}$  grows with the  $2p$ -th power of the distance between the actual argument of measurement  $\chi^{i\flat}$  and the position of interest  $\xi$ .

Now the main step to solving Eq. 14 is to see that the posterior of  $\boldsymbol{\varphi}$  given  $\mathbf{x}, \mathbf{y}$ , and  $\boldsymbol{\gamma}$  is still Gaussian,

$$\varrho(\boldsymbol{\varphi}|\mathbf{x}, \mathbf{y}, \boldsymbol{\gamma}) \propto \exp\{-(\boldsymbol{\varphi} - \tilde{\boldsymbol{\mu}}_\varphi)' \tilde{P}_\varphi (\boldsymbol{\varphi} - \tilde{\boldsymbol{\mu}}_\varphi)/2\}, \quad (19)$$

with precision matrix  $\tilde{P}_\varphi = \tilde{P}_\varphi(\boldsymbol{\gamma})$  and mean  $\tilde{\boldsymbol{\mu}}_\varphi = \tilde{\boldsymbol{\mu}}_\varphi(\boldsymbol{\gamma})$  given by

$$\boxed{\tilde{P}_\varphi(\boldsymbol{\gamma}) = P_\varphi + X'WX, \quad \tilde{\boldsymbol{\mu}}_\varphi(\boldsymbol{\gamma}) = \tilde{P}_\varphi^{-1} \{P_\varphi \boldsymbol{\mu}_\varphi + X'W(\mathbf{y} - \boldsymbol{\mu}_r)\}}, \quad (20)$$

where we call

$$W = (\Sigma_r + \Sigma_\varepsilon)^{-1} \quad (21)$$

the *squared weight matrix*. Note that the basic behaviour of  $W^{i,j\flat}$  is to decrease with a growing distance of  $\chi^{i\flat}$  and  $\chi^{j\flat}$  from  $\xi$ , similar to the weights used in IDW.

*Singularities.* Although in pathological cases, some of the involved matrices can be singular, they are nonsingular in general. More precisely, assume that  $\boldsymbol{\gamma}$  has a continuous distribution and  $\Sigma_\psi$  is nonsingular, so that  $\Sigma_r$  is nonsingular, too (see Eq. 18). Then  $W$  exist and is nonsingular for almost all choices of  $\Sigma_\varepsilon$ , including the case  $\Sigma_\varepsilon = 0$ . In that case, and for generic  $\chi$  and  $\xi$ ,  $X$  has full rank  $\min(N, m)$  (like a Vandermonde matrix), so that also  $X'WX$  is nonsingular when  $m \leq N$ . Finally, under these assumptions  $\tilde{P}_\varphi$  is nonsingular for almost all choices of  $P_\varphi$ , including the case  $P_\varphi = 0$ .

### 3.3. The Moving Taylor Bayesian Regression estimator

Integrating over all possible values of the argument errors  $\gamma$  finally shows that the posterior of  $\varphi$  given  $\mathbf{x}$  and  $\mathbf{y}$  is a mixture of Gaussians whose mean and covariance matrix are

$$\hat{\boldsymbol{\varphi}} = \int d^N \gamma \varrho(\gamma|\mathbf{x}) \tilde{\boldsymbol{\mu}}_\varphi(\gamma), \quad (22)$$

$$\hat{\Sigma}_\varphi = \int d^N \gamma \varrho(\gamma|\mathbf{x}) \left\{ \tilde{P}_\varphi(\gamma)^{-1} + (\tilde{\boldsymbol{\mu}}_\varphi(\gamma) - \hat{\boldsymbol{\varphi}})(\tilde{\boldsymbol{\mu}}_\varphi(\gamma) - \hat{\boldsymbol{\varphi}})' \right\}. \quad (23)$$

This posterior mean, which we call the *MOTABAR estimator*, can now be used as a natural estimate of  $\varphi$  given  $\mathbf{x}$ ,  $\mathbf{y}$ , and one can see that this estimate is

- an affinely linear function of the measured values  $y^{(i)}$ , with weights that decrease with growing distance between  $\xi$  and  $x^{(i)}$ , and
- a mixture of values  $\tilde{\boldsymbol{\mu}}_\varphi(\gamma)$  each of which is an infinitely smooth rational function of the position of interest  $\xi$  that has no poles.

Because the entries of the inverse of an  $m \times m$  square matrix  $A$  are rational functions of the entries of  $A$  of degree at most  $(m-1, m)$ , the degree of  $\tilde{\boldsymbol{\mu}}_\varphi(\gamma)$  is at most  $(2p^2N, 2p^2N)$ . In practice, one can determine  $\tilde{\boldsymbol{\mu}}_\varphi(\gamma)$  and  $\tilde{P}_\varphi(\gamma)$  like this, avoiding the large  $N \times N$  matrix inversions and solving instead two systems of linear equations:

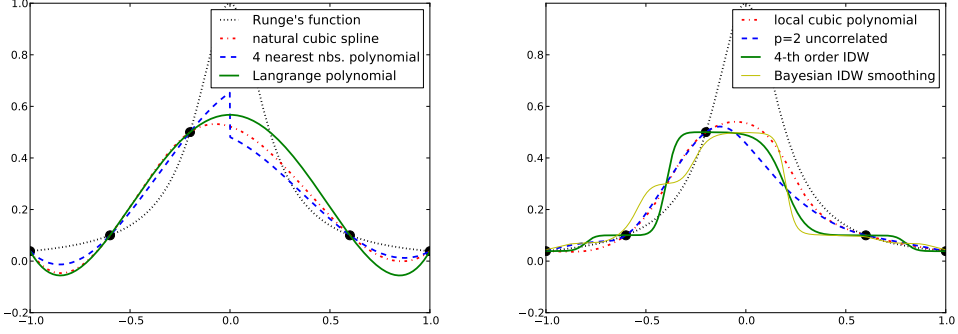
- Given  $\xi$ ,  $\mathbf{x}$ , and  $\gamma$ , compute  $X$ ,  $\boldsymbol{\mu}_r$ , and  $\Sigma_r$  from Eqns. 12 and 7.
- Determine  $A = WX$  and  $B = W(\mathbf{y} - \boldsymbol{\mu}_r)$  by solving  $(\Sigma_r + \Sigma_\varepsilon)(A, B) = (X, \mathbf{y} - \boldsymbol{\mu}_r)$ .
- Compute  $\tilde{P}_\varphi = P_\varphi + X'A$ .
- Determine  $\tilde{\boldsymbol{\mu}}_\varphi$  by solving  $\tilde{P}_\varphi \tilde{\boldsymbol{\mu}}_\varphi = P_\varphi \boldsymbol{\mu}_\varphi + X'B$ .

Note that in pathological cases, either of the two systems might not be solvable uniquely, which possibility we do not discuss here. Although the usually smaller  $m \times m$  matrix inversions needed to determine  $\hat{\Sigma}_\varphi$  from in Eq. 23 are not as easily avoided, one can at least determine the posterior variance of  $f(\xi) = \varphi^{(0)}$  by solving  $\tilde{P}_\varphi \mathbf{c} = (1, 0, \dots, 0)'$  for  $\mathbf{c}$  and using  $(\tilde{P}_\varphi^{-1})^{(0,0)} = c^{(0)}$  in the integral in Eq. 23 to determine  $\hat{\Sigma}_\varphi^{(0,0)}$ . If the argument errors  $\gamma$  are not known to be zero, the integrals over  $\gamma$  will usually have to be evaluated numerically even when  $\gamma$  is Gaussian as well, since the integrand is a nonlinear function of  $\gamma$ . If  $\gamma$  has small variance, it can however be feasible to approximate the integrals by using a quadratic or 4th-order approximation of  $\tilde{\boldsymbol{\mu}}_\varphi(\gamma)$ , which will be explored in a separate article.

### 3.4. Equivariance properties

In addition to the obvious translation invariance in all dimensions, the MOTABAR estimator is equivariant under various linear scaling transformations: (i) When  $\Sigma_\varepsilon$ ,  $\Sigma_\varphi$ , and  $\Sigma_\psi$  are all multiplied with the same constant  $C > 0$ , then  $\hat{\boldsymbol{\varphi}}$  remains unchanged and  $\hat{\Sigma}_\varphi$  is multiplied by  $C$  as well. (ii) When for all  $i$ , all  $\alpha$ , and some  $k$ ,  $\xi_k$ ,  $x_k^{(i)}$ , and  $(\mu_\gamma^{(i)})_k$  are all multiplied with the same constant  $C > 0$ , the distribution of  $\gamma$  is stretched by a factor of  $C$  along the  $k$ -th axis, and when  $\mu_\varphi^{(\alpha)}$ ,  $\mu_\psi^{(\alpha)}$ , the  $\alpha$ -th row and column of  $\Sigma_\varphi$ , and the  $\alpha$ -th row and column of  $\Sigma_\psi$  are all divided by  $C^{\alpha_k}$ , then also  $\hat{\boldsymbol{\varphi}}^{(\alpha)}$  and the  $\alpha$ -th row and column of  $\hat{\Sigma}_\varphi$  are divided by  $C^{\alpha_k}$  for all  $\alpha$ .





**Fig. 2.** (Colour online) Illustration of special and limit cases of MOTABAR interpolation for an equidistant sample (black dots) from Runge's function  $1/(1 + 25x^2)$  with one missing measurement at 0.2. Left: Spline interpolation compared to MOTABAR cases 4.13 (piecewise polynomial based on four nearest neighbours) and 4.8 (Lagrange polynomial). Right: MOTABAR cases 4.4 (local cubic polynomial regression interpolation), 4.11 ( $p = 2$  with uncorrelated errors and priors), 4.12 (4-th order IDW), and a smoothing case ( $p = 5$ , small error, penalised intermediate derivatives), all resulting in nonpolynomial rational functions.

#### 4. Special and limit cases

To better understand the effects of the various inputs to MOTABAR, let us consider a number of special and limit cases, some of which turn out to be equivalent to well-known existing methods, whereas others are presented to show limitations of our approach. Fig. 2 illustrates the diversity of results one can get from the same data.

##### 4.1. Vanishing argument errors

For vanishing argument errors,  $\varrho(\gamma|x)$  becomes a Dirac delta function, so that

$$\hat{\varphi} = \tilde{\mu}_\varphi(\gamma = 0), \quad \hat{\Sigma}_\varphi = \tilde{P}_\varphi(\gamma = 0)^{-1}/2, \quad (24)$$

hence the estimate is a pole-free rational function of  $\xi$  of degree at most  $(2p^2N, 2p^2N)$ . For most of the remainder of this section, we only derive  $\tilde{P}_\varphi = \tilde{P}_\varphi(\gamma)$  and  $\tilde{\mu}_\varphi = \tilde{\mu}_\varphi(\gamma)$  which can then be plugged into Eqs. 22 and 23 if argument errors exist.

##### 4.2. Noninformative priors

If there is no prior information about  $f$ , an improper  $\varphi$ -prior with  $P_\varphi \rightarrow 0$  and a centralised  $\psi$ -prior with  $\mu_\psi \rightarrow 0$  can be used, in which case

$$\tilde{P}_\varphi \rightarrow X'WX, \quad \tilde{\mu}_\varphi \rightarrow (X'WX)^{-1}X'Wy. \quad (25)$$

In this case,  $\tilde{\mu}_\varphi$  is that value of  $\varphi$  which minimises the quadratic function

$$L(\varphi) = \|A\varphi - b\|_2^2 \quad \text{with} \quad A = W^{1/2}X, \quad b = W^{1/2}y, \quad (26)$$

which provides an alternative way of numerical computation. Although this is formally a weighted least-squares estimator, it does not estimate a global set of parameters  $\beta$  as in a global linear regression model  $y = X\beta + \varepsilon$ , but is a local model in which  $X$  and  $W$  depend on  $\xi$ , so that the terms in the estimator have to be computed for each  $\xi$  separately.

*Consistent estimation of polynomial components.* If  $P_\varphi = 0$  and  $\boldsymbol{\mu}_\psi = 0$ , we have  $\tilde{P}_\varphi^{-1} X' W X = I$ . Hence if  $\xi = 0$  and  $\mathbf{y}$  is a monomial with  $y^{(i)} = (x^{(i)})^\alpha / \alpha!$  for some  $\alpha$  with  $|\alpha| < p$ , then  $\mathbf{y}$  equals the  $\alpha$ -th column of  $X$  and hence  $\tilde{\boldsymbol{\mu}}_\varphi = \tilde{P}_\varphi^{-1} X' W \mathbf{y}$  equals the  $\alpha$ -th column of  $I$ . This means that the estimate  $\tilde{\boldsymbol{\mu}}_\varphi^{(\beta)}$  is zero for  $\alpha \neq \beta$  and one for  $\alpha = \beta$ , which are the correct derivatives of the given monomial at  $\xi = 0$ . Consequently, MOTABAR with improper priors is also equivariant under the addition of polynomials, e. g., a quadratic trend in a time series, a property shared with numerical differentiation via discrete Legendre polynomials. However, while the latter has  $\hat{\varphi} = C(\mathbf{x}, \xi) \mathbf{y}$  with an orthogonal coefficient matrix  $C(\mathbf{x}, \xi)$ , the MOTABAR coefficient matrix  $\tilde{P}_\varphi^{-1} X' W$  is not orthogonal in general.

For proper priors with  $P_\varphi \neq 0$  and  $\boldsymbol{\mu}_\varphi = 0$ , the absolute value of the  $\alpha$ -th derivative is underestimated since the prior drags the estimate towards its mean at zero, and the absolute value of the other derivatives is slightly overestimated (nonzero). This effect becomes smaller with increasing  $p$ .

#### 4.3. Vanishing value errors: interpolation

For vanishing value errors, we have  $\Sigma_\varepsilon \rightarrow 0$ , hence  $W \rightarrow P_r = \Sigma_r^{-1}$  and

$$\tilde{P}_\varphi \rightarrow P_\varphi + X' P_r X, \quad \tilde{\boldsymbol{\mu}}_\varphi \rightarrow \tilde{P}_\varphi^{-1} \{P_\varphi \boldsymbol{\mu}_\varphi + X' P_r (\mathbf{y} - \boldsymbol{\mu}_r)\}. \quad (27)$$

In the limit case of  $\Sigma_\varepsilon = 0$ , this can also be derived directly from Eq. 14 by substituting  $\mathbf{r} = \mathbf{y} - X\boldsymbol{\varphi}$ . In that case, however,  $\Sigma_r$  can become singular when  $\xi \rightarrow \chi^{(i)}$  for some  $i$ , which has to be taken care of in the numerical solution, e. g., by assuming very small but nonzero  $\Sigma_\varepsilon$ , or by using the following exact solution for the case  $\xi = \chi^{(i)}$ : In that case,  $\varphi^{(0)} = y^{(i)}$  and, using the notation  $M_{(j,k)}$ ,  $\mathbf{v}_{(j)}$  for a matrix  $M$  without its  $j$ th row and  $k$ th column and a vector  $\mathbf{v}$  without its  $j$ th element,

$$\tilde{P}_{\varphi,(0,0)} = Q + X'_{(i,0)} \Sigma_{r,(i,i)}^{-1} X_{(i,0)}, \quad (28)$$

$$\tilde{\boldsymbol{\mu}}_{\varphi,(0)} = \tilde{P}_{\varphi,(0,0)}^{-1} \{Q\mathbf{v} + X'_{(i,0)} (\Sigma_{r,(i,i)})^{-1} (\mathbf{y}_{(i)} - y^{(i)} \mathbf{e} - \boldsymbol{\mu}_{r,(i)})\}, \quad (29)$$

where  $Q, \mathbf{v}$  are the prior precision and mean of  $\boldsymbol{\varphi}_{(0)}$  conditional on  $\varphi^{(0)} = y^{(i)}$ , and  $\mathbf{e} = (1, \dots, 1)'$  is a vector of ones.

#### 4.4. Vanishing value errors with noninformative priors: local polynomial regression interpolation

If in addition to  $\Sigma_\varepsilon = 0$  we have  $P_\varphi = 0$  and  $\boldsymbol{\mu}_\psi = 0$ , we get

$$\tilde{P}_\varphi = X' P_r X, \quad \tilde{\boldsymbol{\mu}}_\varphi = (X' P_r X)^{-1} X' P_r \mathbf{y} \quad (30)$$

for general  $\xi$ , while for  $\xi = \chi^{(i)}$ , we get  $\varphi^{(0)} = y^{(i)}$ ,  $\tilde{P}_{\varphi,(0,0)} = X'_{(i,0)} V X_{(i,0)}$ , and

$$\tilde{\boldsymbol{\mu}}_{\varphi,(0)} = (X'_{(i,0)} V X_{(i,0)})^{-1} X'_{(i,0)} V (\mathbf{y}_{(i)} - y^{(i)} \mathbf{e}) \quad (31)$$

with  $V = \Sigma_{r,(i,i)}^{-1}$ . This is interpolation based on local polynomial regression in which the weights  $P_r^{(i,j)}$  decrease as a power of a quadratic form of  $\chi^{(i)} - \xi$  and  $\chi^{(j)} - \xi$  (e. g., the dash-dotted red line in Fig. 2 right). Hence, in contrast to other local polynomial methods such as LOESS, far away observations get a positive though small weight.

#### 4.5. Diverging errors

At the other end of the error scale, for  $P_\varepsilon \rightarrow 0$ , we get  $W \rightarrow 0$  and thus  $\tilde{P}_\varphi \rightarrow P_\varphi$  and  $\tilde{\mu}_\varphi \rightarrow \mu_\varphi$ . So if the  $\varphi$ -prior is proper, the posterior approximates it, otherwise it diverges. The same behaviour obtains for a decreasingly informative  $\psi$ -prior, i.e., for  $P_\psi \rightarrow 0$ . This shows that while a noninformative  $\varphi$ -prior might be chosen, the  $\psi$ -prior must be proper since it regulates the overall variability of the estimate.

#### 4.6. Penalised higher derivatives

If the value error variance is finite ( $\Sigma_\varepsilon \neq 0$ ), one might, on the other hand, assume that the derivatives of  $f$  of degree  $\geq p$  are negligible compared to the scale of  $\varepsilon$ , e.g., because  $f$  is believed to be approximately a polynomial of degree at most  $p - 1$ . Then, taking the limit  $\mu_\psi \rightarrow 0$  and  $\Sigma_\psi \rightarrow 0$ , we get

$$\tilde{P}_\varphi \rightarrow P_\varphi + X'P_\varepsilon X, \quad \tilde{\mu}_\varphi \rightarrow \tilde{P}_\varphi^{-1} (P_\varphi \mu_\varphi + X'P_\varepsilon y), \quad (32)$$

which is a form of Bayesian polynomial regression. Although the squared weight matrix  $W = P_\varepsilon$  no longer depends on  $\xi$  here, this is still not a global regression method since the  $\varphi$ -prior might depend on  $\xi$ , e.g., because one assumes some linear or nonlinear trend or some change in variability with  $\xi$ .

#### 4.7. Penalised higher derivatives with a noninformative prior: ordinary global polynomial regression

A truly global regression method can be obtained by using the noninformative  $\varphi$ -prior with  $P_\varphi \rightarrow 0$  in addition to  $\mu_\psi \rightarrow 0$  and  $\Sigma_\psi \rightarrow 0$ . If  $N \geq \binom{p-1+d}{d}$ , then

$$\tilde{P}_\varphi \rightarrow X'P_\varepsilon X, \quad \tilde{\mu}_\varphi \rightarrow (X'P_\varepsilon X)^{-1} X'P_\varepsilon y. \quad (33)$$

In the limit, this is just ordinary global least-squares polynomial regression with polynomials of order  $p - 1$  and possibly correlated value errors of differing magnitude.

##### 4.7.1. Linear regression with argument errors vs. total least squares

For  $d = 1$ ,  $p = 2$ ,  $\varphi = (a, b)'$ , and nonvanishing argument measurement errors, our model becomes the “errors in the variables” linear regression model

$$y^{(b)} = a + b(x^{(b)} - \gamma^{(b)}) + \varepsilon^{(b)}. \quad (34)$$

Since this is linear in  $\gamma_i$ , the integral over  $\gamma$  in Eq. 14 can be solved analytically. For  $\xi = 0$ ,  $P_\varphi = 0$ ,  $\mu_\psi = 0$ ,  $\Sigma_\psi = 0$ ,  $\Sigma_\varepsilon = \sigma_\varepsilon^2 I$ , and  $\Sigma_\gamma = \sigma_\gamma^2 I$ , this results in the non-Gaussian posterior

$$\begin{aligned} \varrho(a, b | x, y) &\propto \int d^N \gamma \exp \left( - \|\gamma\|_2^2 / 2\sigma_\gamma^2 - \|\mathbf{y} - a\mathbf{e} - b\mathbf{x} + b\gamma\|_2^2 / 2\sigma_\varepsilon^2 \right) \\ &\propto (\sigma_\varepsilon^2 + b^2 \sigma_\gamma^2)^{-N/2} \exp \left\{ - \|\mathbf{y} - a\mathbf{e} - b\mathbf{x}\|_2^2 / 2(\sigma_\varepsilon^2 + b^2 \sigma_\gamma^2) \right\}. \end{aligned} \quad (35)$$

The posterior mode has  $a = \bar{y} - b\bar{x}$  and

$$N\sigma_\gamma^4 b^3 + \sigma_\gamma^2 s_{xy} b^2 + (N\sigma_\varepsilon^2 \sigma_\gamma^2 + \sigma_\varepsilon^2 s_{xx} - \sigma_\gamma^2 s_{yy})b - \sigma_\varepsilon^2 s_{xy} = 0, \quad (36)$$

where  $\bar{x} = \sum_i x^{i\flat}/N$  and  $\bar{y} = \sum_i y^{i\flat}/N$ ,  $s_{xx} = \sum_i (x^{i\flat} - \bar{x})^2$ ,  $s_{yy} = \sum_i (y^{i\flat} - \bar{y})^2$ , and  $s_{xy} = \sum_i (x^{i\flat} - \bar{x})(y^{i\flat} - \bar{y})$ . If  $s_{xy} > 0$ ,  $b$  is the largest solution of the above equation, otherwise the smallest, and it has the same sign as  $s_{xy}$ .

When the same model is estimated with the *total least-squares* method (aka *Deming regression*) studied already by [Kummell \(1879\)](#), the equation is

$$\sigma_\gamma^2 s_{xy} b^2 + (\sigma_\varepsilon^2 s_{xx} - \sigma_\gamma^2 s_{yy}) b - \sigma_\varepsilon^2 s_{xy} = 0,$$

instead, which is symmetric under an exchange of “dependent” and “independent” variables and leads to a larger absolute value of  $b$ . For  $\sigma_\gamma^2 \ll s_{xx}/N$ , this difference vanishes, and for  $\sigma_\gamma \rightarrow 0$ , both solutions converge to the ordinary least-squares linear regression line with  $b \rightarrow s_{xy}/s_{xx}$ . For  $\sigma_\varepsilon \rightarrow 0$ , total least squares gives  $b \rightarrow s_{yy}/s_{xy}$  independently of  $\sigma_\gamma$ , while the cubic equation gives

$$b \rightarrow \left\{ \text{sign}(s_{xy}) \sqrt{4N\sigma_\gamma^2 s_{yy} + s_{xy}^2 - s_{xy}} \right\} / 2N\sigma_\gamma^2,$$

which still depends on  $\sigma_\gamma$  and has  $b \approx \text{sign}(s_{xy}) \sqrt{s_{yy}/N}/\sigma_\gamma \rightarrow 0$  for  $\sigma_\gamma \rightarrow \infty$ . This comparison shows that unlike in total least squares, it is essential in MOTABAR whether we consider  $y$  a function of  $x$  or vice versa.

To understand why the effects of value and argument errors are different in the linear model and why the cubic equation is not symmetric in  $x, y$ , consider the simple case where the measurements are  $\mathbf{x} = \mathbf{y} = (-2, 0, 2)$  and only the middle one,  $(x^{2\flat}, y^{2\flat}) = (0, 0)$ , has errors in both dimensions which are independent with  $\gamma^{2\flat}, \varepsilon^{2\flat} \in \{1, -1\}$  with equal probability. Then the real data  $\chi, f(\chi)$  are either  $(-2, -1, 2), (-2, -1, 2)$  or  $(-2, 1, 2), (-2, 1, 2)$ , both giving slope  $b = 1$  in linear regression, or  $(-2, -1, 2), (-2, 1, 2)$  or  $(-2, 1, 2), (-2, -1, 2)$ , both giving slope  $b \approx 0.84$ , so that the MOTABAR posterior mean is  $b \approx 0.92$ , whereas total least squares results in  $b = 1$  because of the obvious symmetry.

#### 4.8. Larger order with non-informative prior: Lagrange interpolation

For  $d = 1$ ,  $N = m = p$ , and  $P_\varphi = 0$ , we get Lagrange interpolation with  $\tilde{\mu}_\varphi = X^{-1}\mathbf{y}$  (e. g., the solid green line in Fig. 2 left). For  $d > 1$  and  $N = m$ , it depends on the placement of the  $\chi^{i\flat}$  whether the resulting polynomial is an interpolant or not (see [Gasca and Sauer \(2000\)](#) for an overview).

#### 4.9. Minimal order with uncorrelated errors and priors: Bayesian IDW smoothing of order two

A particularly simple case occurs for the minimal choice of  $p$ ,  $p = 1$ , and when both value errors and priors are uncorrelated, so that  $\Sigma_\varepsilon$  and  $\Sigma_\psi$  are diagonal, and  $\Sigma_\varphi = (\sigma_\varphi^2)$ . Then we have  $\boldsymbol{\varphi} = (f(\xi))$ ,  $X = (1, \dots, 1)'$ , and  $W^{i,j} = \delta_{ij} w^{i\flat}$ , where

$$w^{i\flat} = \frac{1}{s^{i\flat} + \Sigma_\varepsilon^{i,i}}, \quad s^{i\flat} = \sum_{k=1}^d \Sigma_\psi^{ki,ki} (\chi_k^{i\flat} - \xi_k)^2. \quad (37)$$

The latter is the squared distance between  $\chi^{i\flat}$  and  $\xi$  as measured by the quadratic norm that is weighted with the prior derivative variances  $\Sigma_\psi^{k,k}$ . The posterior distribution of  $f(\xi)$  given  $\boldsymbol{\gamma}$  is then Gaussian with mean and variance given by

$$\tilde{\mu}_\varphi = \frac{\mu_\varphi + \sigma_\varphi^2 \sum_i w^{i\flat} (y^{i\flat} - \mu_r^{i\flat})}{1 + \sigma_\varphi^2 \sum_i w^{i\flat}}, \quad \tilde{\sigma}_\varphi^2 = \frac{\sigma_\varphi^2}{1 + \sigma_\varphi^2 \sum_i w^{i\flat}}. \quad (38)$$

This is a generalization of inverse distance weighting that takes into account value error via the occurrence of  $\sigma_\varepsilon^{\zeta^i}$  in Eq. 37 and prior information via the occurrence of  $\mu_\varphi, \sigma_\varphi^2$  in Eq. 38. We propose to call this method *Bayesian IDW smoothing*. Note that, as expected,  $\tilde{\mu}_\varphi$  is a rational function of  $\xi$  of degree at most  $(2p^2N, 2p^2N) = (2N, 2N)$ .

#### 4.10. Minimal order with vanishing errors and noninformative priors: ordinary IDW with squared distances

Setting  $\Sigma_\varepsilon = 0$ ,  $P_\varphi = 0$ , and  $\Sigma_\psi = \sigma_\psi^2 I$  in the preceding, we get ordinary IDW with squared distances,

$$w^{\zeta^i} = \frac{1}{\sigma_\psi^2 \|\chi_k^{\zeta^i} - \xi_k\|_2^2}, \quad \tilde{\mu}_\varphi = \frac{\sum_i w^{\zeta^i} y^{\zeta^i}}{\sum_i w^{\zeta^i}}, \quad \tilde{\sigma}_\varphi^2 = \frac{1}{\sum_i w^{\zeta^i}}, \quad (39)$$

where  $\tilde{\mu}_\varphi$  is now a rational function of  $\xi$  of degree exactly  $(2N - 2, 2N - 2)$ . Note that while  $\tilde{\mu}_\varphi$  is independent from  $\sigma_\psi^2$ , the posterior variance is proportional to  $\sigma_\psi^2$ . In other words, our Bayesian approach shows that the uncertainty of the IDW estimate of order two is directly proportional to the scale of the derivatives of  $f$ .

#### 4.11. Order two in one dimension with uncorrelated errors and priors

If  $p = 2$ ,  $d = 1$ ,  $\mu_\varphi = 0$ ,  $\mu_\psi = 0$ ,  $\Sigma_\varepsilon$  and  $\Sigma_\psi$  are diagonal, and  $P_\varphi = \text{diag}(a, b)$ , then

$$\tilde{\mu}_\varphi = \frac{\left( \frac{\sum_i z^{\zeta^i} y^{\zeta^i}}{\sum_i w^{\zeta^i} \{a(\chi^{\zeta^i} - \xi) - \sum_j w^{\zeta^j} (\chi^{\zeta^j} - \chi^{\zeta^i})\} y^{\zeta^i}} \right)}{a\{b + \sum_i w^{\zeta^i} (\chi^{\zeta^i} - \xi)^2\} + \sum_i z^{\zeta^i}}, \quad (40)$$

$$\tilde{P}_\varphi = \begin{pmatrix} a + \sum_i w^{\zeta^i} & \sum_i w^{\zeta^i} (\chi^{\zeta^i} - \xi) \\ \sum_i w^{\zeta^i} (\chi^{\zeta^i} - \xi) & b + \sum_i w^{\zeta^i} (\chi^{\zeta^i} - \xi)^2 \end{pmatrix}, \quad (41)$$

$$z^{\zeta^i} = w^{\zeta^i} \{b + \sum_j w^{\zeta^j} (\chi^{\zeta^j} - \xi)(\chi^{\zeta^j} - \chi^{\zeta^i})\}, \quad (42)$$

$$w^{\zeta^i} = \frac{1}{\Sigma_\psi^{i,i,b} (\chi^{\zeta^i} - \xi)^4 / 2 + \Sigma_\varepsilon^{i,i,b}}. \quad (43)$$

Note that  $\tilde{\mu}_\varphi^{(0)}$  does not only depend on the distances of the  $\chi^{\zeta^i}$  from  $\xi$  as encoded in  $w^{\zeta^i}$ , but also on the relative position of  $\chi^{\zeta^i}$  and  $\chi^{\zeta^j}$ , via the mixture terms in  $z^{\zeta^i}$ . See the dashed blue line in Fig. 2 right for an example. This dependency vanishes if we let  $a \rightarrow 0$  and  $b \rightarrow \infty$ , where we get an instance of the following case:

#### 4.12. Uncorrelated errors, penalised intermediate derivatives, and specific priors: IDW smoothing

Assume that  $\mu_\varphi = 0$ ,  $\mu_\psi = 0$ ,  $\Sigma_\varepsilon$  is diagonal,  $P_\varphi = \text{diag}(a, b, \dots, b)$ , and  $\Sigma_\psi^{\alpha i, \beta j} = \delta_{\alpha\beta} \delta_{ij} v^{\zeta^i}$ . If we let the prior for  $\varphi^{(0)}$  become noninformative by letting  $a \rightarrow 0$ , but let the prior for  $\varphi^{(\alpha)}$  with  $|\alpha| > 0$  become sharp at  $\varphi^{(\alpha)} = 0$  by letting  $b \rightarrow \infty$ , then

$$\tilde{\mu}_\varphi^0 \rightarrow \frac{\sum_i w^{\zeta^i} y^{\zeta^i}}{\sum_i w^{\zeta^i}} \quad \text{with} \quad w^{\zeta^i} = \frac{1}{v^{\zeta^i} \|\chi^{\zeta^i} - \xi\|_2^{2p} / p! + \Sigma_\varepsilon^{i,i,b}}. \quad (44)$$

In other words, the MOTABAR estimate converges to the ordinary IDW smoothing (or interpolation, if  $\Sigma_\varepsilon = 0$ ) solution with exponent  $2p$ . E. g.,  $p = 2$  and  $\Sigma_\varepsilon = 0$  gives 4-th IDW interpolation (solid

green line in Fig. 2 right). Note that this, however, also shows that using IDW with higher powers than two corresponds to implicitly assuming that some derivatives vanish which the result shows not to vanish after all. Hence the only plausible form of IDW is the one with squared distances.

*Inconsistent derivatives.* This example also illustrates an unintuitive feature of MOTABAR: The estimates derivative  $\tilde{\mu}_\varphi^{(\alpha)}$  need not coincide with the derivative of the estimated value  $D^\alpha \tilde{\mu}_\varphi^{(0)}$  with respect to  $\xi$ . In IDW smoothing, the estimated slope is zero as enforced by the sharp prior, although the estimated value is not constant.

#### 4.13. Large order with penalised intermediate derivatives: piecewise estimators

On the upper end of the order spectrum, for certain choices of priors, one can choose very large orders  $p$  without running into numerical infeasibilities. A special limit case obtains with similar priors as above: Assume  $\Sigma_\varepsilon = 0$ , a noninformative prior for the function value  $f(\xi) = \varphi^{(0)}$  (so that  $P_\varphi^{(0,0)} = 0$ ), an uncorrelated and centralised  $\psi$ -prior with  $\mu_\psi = 0$  and  $\Sigma_\psi^{(\alpha i, \beta j)} = \delta_{\alpha\beta} \delta_{ij} \gamma^{i\flat}$ , and that all derivatives of  $f$  of order  $1 \dots p-1$  vanish at  $\xi$  (so that  $\mu_\varphi^{(\alpha)} = 0$  and  $\Sigma_\varphi^{(\alpha, \alpha')} = 0$  for  $|\alpha| > 0$ ). In the limit for  $p \rightarrow \infty$ , then only the nearest neighbour  $\chi^{i\flat}$  of  $\xi$  is used in the estimation. This is because then  $P_r$  is diagonal and  $P_r^{i,j,\flat} / P_r^{i,i,\flat} \rightarrow 0$  for  $j \neq i$ , so that  $\varrho(\varphi^{(0)} | \mathbf{x}, \mathbf{y}, \boldsymbol{\gamma}) \sim \exp\{-P_r^{i,i,\flat} (\varphi^{(0)} - y^{i\flat})^2\}$ . In other words, MOTABAR then estimates  $f(\xi)$  by  $\hat{\varphi}^{(0)} \rightarrow y^{i\flat}$ , as in simple *nearest neighbour estimation*. That is, for large  $p$  and these priors, MOTABAR estimate approximates a step function that is constant on the Voronoi cells (aka Thiessen polyhedrons) of the sample. As the 4-th order IDW example in Fig. 2 (right) shows, the step-like shape can already be seen for small values of  $p$ . A step-like function with a similar smoothening at the edges also obtains for  $p \rightarrow \infty$  when argument errors  $\boldsymbol{\gamma}$  are not vanishing, since then  $\hat{f}(\xi) \rightarrow \sum_i \mathbb{P}(\forall j : \|\chi^{i\flat} - \xi\| \leq \|\chi^{j\flat} - \xi\|) y^{i\flat}$ . On the other hand, nonvanishing value errors, even if very small, can lead to additional levels at the means of more than one nearest neighbour, showing that this limit behaviour is quite unstable (see, e. g., the thin yellow line in Fig. 2 right).

One can also get higher-order piecewise polynomial estimates, by assuming that only the derivatives of  $f$  of order  $q+1 \dots p-1$  vanish at  $\xi$  for some  $q < p$ , and using an improper prior with  $P_\varphi^{(\alpha, \alpha')} = 0$  for  $|\alpha| \leq q$ . Then the MOTABAR estimate approximates a piecewise polynomial interpolant of degree  $q$  that uses the  $\binom{d+q}{d}$  observations nearest to  $\xi$  (assuming those are in general position). For  $q = 1$ , this leads to a piecewise linear estimate that need not, however, coincide with ordinary linear interpolation since, e. g., the two nearest neighbours of  $\xi \in \mathbb{R}$  might both lie left of  $\xi$ , leading to a discontinuity in the estimate to the right of  $\xi$ . Only when the positions  $\chi^{i\flat}$  build a regular polyhedral grid, the estimate approximates ordinary linear interpolation. For  $d = 1$ , this requires equidistant measurement positions, and for  $d = 2$  a regular triangular grid is required, whereas on a square grid, one gets a discontinuous linear interpolation with four square interpolation cells per grid cell, each corresponding to a different choice of three of the four corners of the grid cell. Analogous results hold for higher-dimensional rectangular grids. For  $q = 3$  and  $d = 1$ , the limit result is a piecewise cubic polynomial which is, however, not a cubic spline since although its 2nd derivative is continuous, its value and slope are not (e. g., the dashed blue line in Fig. 2 left).

#### 4.14. Extreme arguments of interest

If the distance between  $\xi$  and the data positions  $\chi^{i\flat}$  diverges, i. e., if  $\|\xi - \chi^{i\flat}\|_2 \rightarrow \infty$ , we have  $X'WX \rightarrow 0$ . If the  $\varphi$ -prior is informative ( $P_\varphi \neq 0$ ), we then have

$$\tilde{P}_\varphi \rightarrow P_\varphi, \quad \tilde{\mu}_\varphi \rightarrow \mu_\varphi, \quad (45)$$

that is, the data are too far away to drag the posterior significantly away from the prior. With a noninformative  $\varphi$ -prior, on the other hand, the estimate for  $\|\xi - \chi^{i\flat}\|_2 \rightarrow \infty$  will either approximate the sample mean (if  $p = 1$ ) or will diverge to  $\pm\infty$  (almost surely for  $p > 1$ ).

#### 4.15. Large amounts of data: conjectured exponential rate of convergence

If the  $\chi^{i\flat}$  are either regularly distributed or drawn from a sufficiently smooth distribution, the distance between  $\xi$  and the closest  $\chi^{i\flat}$  will decay at a rate  $\sim N^{-1/d}$  for  $N \rightarrow \infty$ . We conjecture that then the posterior precision grows at a rate

$$\tilde{P}_\varphi^{\langle\alpha,\beta\rangle} \sim N^{1+(2p-|\alpha|-|\beta|-1)/d} \quad \text{if } \Sigma_\varepsilon = 0 \quad (\text{conjectured}) \quad (46)$$

$$\tilde{P}_\varphi^{\langle\alpha,\beta\rangle} \sim N \quad \text{if } \Sigma_\varepsilon \neq 0 \quad (\text{conjectured}) \quad (47)$$

for nonpathological error distributions and priors. The rationale for this conjecture is this: Assume that  $\Sigma_\varepsilon = 0$ ,  $\Sigma_\psi = \sigma_\psi^2 I$ , and the  $N$  arguments  $\chi^{i\flat}$  are sufficiently uniformly distributed over a  $d$ -dimensional ball of unit radius about  $\xi$ , and let  $|\alpha|, |\beta| < p$ . Then

$$\tilde{P}_\varphi^{\langle\alpha,\beta\rangle} - P_\varphi^{\langle\alpha,\beta\rangle} \propto \sum_{i=1}^N (\chi^{i\flat} - \xi)^{\alpha+\beta} \|\chi^{i\flat} - \xi\|_2^{-2p},$$

and this sum is asymptotically proportional to the integral

$$\begin{aligned} & N \int_{\|x-\xi\|_2 \in [N^{-1/d}, 1]} d^d x (x - \xi)^{\alpha+\beta} \|x - \xi\|_2^{-2p} \\ & \sim N \int_{z=N^{-1/d}}^1 dz z^{|\alpha|+|\beta|} z^{-2p} \sim N^{1+(2p-|\alpha|-|\beta|-1)/d}. \end{aligned}$$

The conjecture implies  $\tilde{\Sigma}_\varphi^{\langle\alpha,\beta\rangle} \sim 1/N^{1+(2p-|\alpha|-|\beta|-1)/d}$ , in particular  $|\hat{\varphi}^{\langle\alpha\rangle} - (D^\alpha f)(\xi)| \sim N^{(1/2+|\alpha|-p)/d-1/2}$ . In particular, for  $d = 1$ , the conjectured rate of convergence of  $\hat{\varphi}^{(0)}$  to  $f(\xi)$  is  $N^p$ , the same as for spline interpolation.

## 5. Comparison with Wang et al.'s approach

Wang et al. (2010a) introduced a different approximation scheme that is also based on a local Taylor expansion. Adapting their terminology and notation to ours, that scheme can be described as defining an estimator

$$\hat{f}_{\text{Wang}}(\xi) = \mathbf{a}' \mathbf{y} \quad (48)$$

for  $f(\xi)$ , where the weights  $a_i$  are chosen to minimise the squared norm

$$\|\mathbf{a}\|_{\text{Wang}}^2 = \mathbf{a}' (X V_\varphi X' + V_y) \mathbf{a} \quad (49)$$

subject to a number of constraints

$$\sum_i a^{i\flat} = 1, \quad \sum_i a^{i\flat} X^{i,\alpha\flat} = 0 \quad (50)$$

for all  $\alpha$  with  $0 < |\alpha| < q$  for some  $q < p$ , where  $V_\varphi$  and  $V_y$  are diagonal matrices whose entries are certain parameters  $\sigma_\varphi^{\langle\alpha\rangle}$  and error variances,

$$V_\varphi^{\langle\alpha,\alpha\rangle} = (\sigma_\varphi^{\langle\alpha\rangle})^2, \quad V_y^{i,\flat} = (\sigma_r^{i\flat})^2 + (\sigma_\varepsilon^{i\flat})^2. \quad (51)$$

Although the parameters  $(\sigma_\varphi^{(a)})^2$  correspond to the prior variances in our approach, Wang et al.'s rationale is not Bayesian. Their motivation is rather that  $\|\mathbf{a}\|_{\text{Wang}}^2$  is an estimator of the squared approximation error  $\{\hat{f}_{\text{Wang}}(\xi) - f(\xi)\}^2$ , and the constraints make sure that the estimator is correct if the true  $f$  is a polynomial of degree  $\leq q$ . As we see from the usage of  $V_\varphi$  and  $V_y$  instead of  $\Sigma_\varphi$  and  $\Sigma_y = W^{-1}$ , their scheme implicitly assumes uncorrelated ‘‘priors’’ and errors, but one can easily adapt it to deal with correlations by using

$$\|\mathbf{a}\|_{\text{corr.}}^2 = \mathbf{a}'(X\Sigma_\varphi X' + \Sigma_y)\mathbf{a} \quad (52)$$

instead of  $\|\mathbf{a}\|_{\text{Wang}}^2$ . Also, they do not consider the case of argument errors, and for these it does not seem obvious how to deal with them consistently in their framework.

To compare their scheme to ours in the case of no argument errors, note that the MOTABAR estimate  $\tilde{\boldsymbol{\mu}}_\varphi$  is a linear function of  $\mathbf{y}$  only when the prior for  $\boldsymbol{\varphi}$  is improper and the prior for  $\boldsymbol{\psi}$  is centralised, while otherwise the estimate is only an *affine* function of  $\mathbf{y}$ . Hence let us assume  $P_\varphi = 0$  and  $\boldsymbol{\mu}_\psi = 0$  in this section. In that case, the MOTABAR estimator  $\tilde{\boldsymbol{\mu}}_\varphi$  itself is that  $\boldsymbol{\varphi}$  which minimises the quadratic function

$$L(\boldsymbol{\varphi}) = \|\mathbf{A}\boldsymbol{\varphi} - \mathbf{b}\|_2^2 \quad \text{with} \quad \mathbf{A} = W^{1/2}X, \quad \mathbf{b} = W^{1/2}\mathbf{y}. \quad (53)$$

However, while in Wang et al.'s approach a constrained problem obtains, the above is an unconstrained problem. Taking the corresponding limit of  $P_\varphi \rightarrow 0$  in Wang et al.'s scheme, their target function becomes

$$\|\mathbf{a}\|_{\text{corr.}}^2 \approx \mathbf{a}'X\Sigma_\varphi X'\mathbf{a} \quad (54)$$

so that their coefficients depend on the relative variances (and correlation structure) of  $\boldsymbol{\varphi}$  but no longer on  $\Sigma_y$ , while ours depends on the latter but not on the former.

Despite these differences, both approaches seem to have the same rate of convergence for  $N \rightarrow \infty$  and reduce to ordinary polynomial regression or IDW for certain choices of control parameters.

## 6. Non-Gaussian cases and constraints

Prior knowledge about  $\boldsymbol{\varphi}$ , such as constraints of the form  $(D^\alpha f)(x) \in [a, b]$  for some  $\alpha$  and  $a, b \in \mathbb{R} \cup \{\pm\infty\}$ , can easily be incorporated into MOTABAR estimation by choosing a suitable prior distribution for  $\boldsymbol{\varphi}$ , such as a uniform distribution of  $\varphi^{(a)}$  on  $[a, b]$  or a Gaussian restricted to this interval. If one then rewrites the chosen prior in the form

$$\varrho(\boldsymbol{\varphi}) \propto \varrho_0(\boldsymbol{\varphi}) \exp\{-(\boldsymbol{\varphi} - \boldsymbol{\mu}_\varphi)'P_\varphi(\boldsymbol{\varphi} - \boldsymbol{\mu}_\varphi)\}, \quad (55)$$

which can be seen as a generalization of Eq. 16, it is easy to see that the resulting posterior simply involves the same factor  $\varrho_0(\boldsymbol{\varphi})$ , leading to a posterior distribution of

$$\varrho(\boldsymbol{\varphi}|\mathbf{x}, \mathbf{y}) \propto \varrho_0(\boldsymbol{\varphi}) \int d\boldsymbol{\gamma} \varrho(\boldsymbol{\gamma}|\mathbf{x}) \exp\{-(\boldsymbol{\varphi} - \tilde{\boldsymbol{\mu}}_\varphi)' \tilde{P}_\varphi(\boldsymbol{\varphi} - \tilde{\boldsymbol{\mu}}_\varphi)\} \quad (56)$$

with the same  $\tilde{\boldsymbol{\mu}}_\varphi$  and  $\tilde{P}_\varphi$  as before. However, since the posterior mean is then no longer equal to  $\tilde{\boldsymbol{\mu}}_\varphi$ , one either needs to integrate Eq. 56 explicitly or use the posterior mode instead.

In some cases, the posterior mean can be determined analytically. E. g., two-sided constraints of the form  $(D^\alpha f)(x) \in [a, b]$  with finite bounds  $a, b \in \mathbb{R}$  can most easily be modelled by putting  $P_\varphi^{(a, a)} = 0$  and  $\varrho_0(\boldsymbol{\varphi}) = \Theta(\varphi^{(a)} - a^{(a)})\Theta(b^{(a)} - \varphi^{(a)})$ , where  $\Theta$  is the Heaviside step function with  $\Theta(\varphi) = 1$  for  $\varphi \geq 0$  and  $\Theta(\varphi) = 0$  otherwise. For one-sided constraints of the form  $(D^\alpha f)(x) \geq a^{(a)}$ , one can use  $\varrho_0(\boldsymbol{\varphi}) = \Theta(\varphi^{(a)} - a^{(a)})$ .



### 6.1. Function value constrained to an interval

Using the prior factor  $\varrho_0(\boldsymbol{\varphi}) = \Theta(\varphi^{(0)} - a)\Theta(b - \varphi^{(0)})$ , the marginal posterior density of  $\varphi^{(0)}$  is proportional to

$$\Theta(\varphi^{(0)} - a)\Theta(b - \varphi^{(0)}) \exp\{-(\varphi^{(0)} - \mu)^2/2\sigma^2\} \quad (57)$$

with  $\mu = \tilde{\mu}_\varphi^{(0)}$  and  $\sigma^2 = (\tilde{P}_\varphi^{-1}/2)^{(0,0)}$ . For a density of this form, the mean is given by

$$\nu = \mu + \sigma \frac{\exp\{-(\mu - a)^2/2\sigma^2\} - \exp\{-(\mu - b)^2/2\sigma^2\}}{\operatorname{erf}\{(\mu - a)/\sqrt{2}\sigma\} - \operatorname{erf}\{(\mu - b)/\sqrt{2}\sigma\}} \sqrt{\frac{2}{\pi}} \in (a, b). \quad (58)$$

For  $\mu$  outside  $[a, b]$ , this asymptotically equals  $b - \sigma(-1/z + 2/z^3 - 10/z^5 + 74/z^7 - 706/z^9)$  with  $z = (b - \mu)/\sigma$  for  $\mu \rightarrow \infty$  (good approximation for  $\mu > b + 4\sigma$ ), or  $a + \sigma(-1/z + 2/z^3 - 10/z^5 + 74/z^7 - 706/z^9)$  with  $z = (\mu - a)/\sigma$  for  $\mu \rightarrow -\infty$  (good approximation for  $\mu < a - 4\sigma$ ). With this constraint, the MOTABAR estimator becomes

$$\hat{\boldsymbol{\varphi}} = \int d\boldsymbol{\gamma} \varrho(\boldsymbol{\gamma}|\mathbf{x}) \boldsymbol{\nu}, \quad (59)$$

where  $\boldsymbol{\nu}$  depends on  $\boldsymbol{\gamma}$  via  $\mu$  and  $\sigma$ .

### 6.2. One-sided constraints

Letting  $b \rightarrow \infty$  in the preceding, we get

$$\nu = \mu + \sigma \frac{\exp\{-(\mu - a)^2/2\sigma^2\}}{\operatorname{erf}\{(\mu - a)/\sqrt{2}\sigma\} + 1} \sqrt{\frac{2}{\pi}} > a. \quad (60)$$

### 6.3. Non-Gaussian errors and quantile regression

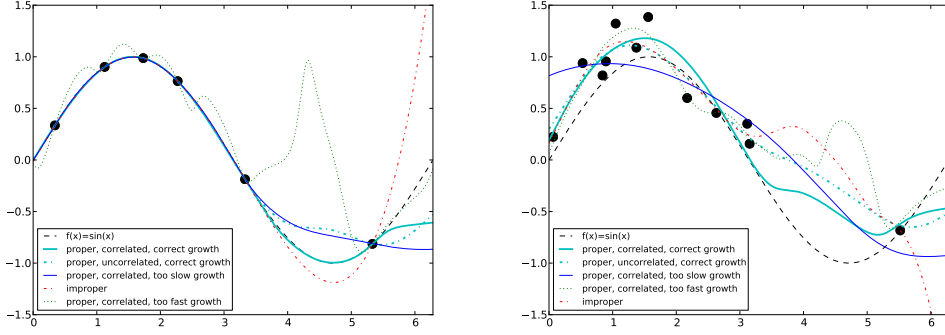
To see that our methodology can be fruitful also in the case of non-Gaussian  $\psi$ -priors and value errors, consider the case of  $d = 1$ , vanishing argument errors, an improper  $\varphi$ -prior, independent Laplace  $\psi$ -priors with  $\varrho(\psi^{(i)}) \propto \exp(-b^{(i)}|\psi^{(i)}|)$  for some  $b^{(i)} > 0$ , and independent value errors with asymmetric Laplace distributions  $\varrho(\varepsilon^{(i)}) \propto \exp\{-a^{(i)}\varrho_\tau(\varepsilon^{(i)})\}$ , where  $\varrho_\tau(z) = z(\tau - \Theta(-z))$ ,  $a^{(i)} > 0$ , and  $\tau^{(i)} \in [0, 1]$ . Then

$$\begin{aligned} \varrho(\mathbf{y}|\mathbf{x}, \boldsymbol{\varphi}) &\propto \prod_i \int dr^{(i)} \exp\{-a^{(i)}\varrho_{\tau^{(i)}}(y^{(i)} - (X\boldsymbol{\varphi})^{(i)} - r^{(i)}) - b^{(i)}| \frac{r^{(i)}}{X^{(i,p)}} |\} \\ &\propto \prod_i g(y^{(i)} - (X\boldsymbol{\varphi})^{(i)}, a^{(i)}, \tau^{(i)}, b^{(i)} / |X^{(i,p)}|), \end{aligned} \quad (61)$$

where

$$g(z, a, \tau, c) = \begin{cases} \frac{\exp\{za(1-\tau)\}}{c+a(1-\tau)} + \frac{\exp\{za(1-\tau)\}-\exp(zc)}{c-a(1-\tau)} + \frac{\exp(zc)}{c+a\tau} & \text{if } z \leq 0 \\ \frac{\exp(-za\tau)}{c+a\tau} + \frac{\exp(-za\tau)-\exp(-zc)}{c-a\tau} + \frac{\exp(-zc)}{c+a(1-\tau)} & \text{if } z \geq 0 \end{cases}, \quad (62)$$

and the posterior mode of  $\boldsymbol{\varphi}$  given  $\mathbf{x}$ ,  $\mathbf{y}$  is the one that maximises  $\varrho(\mathbf{y}|\mathbf{x}, \boldsymbol{\varphi})$ . For  $a^{(i)} \equiv a$ ,  $\tau^{(i)} \equiv \tau$ , and sharp  $\psi$ -priors with  $b^{(i)} \rightarrow \infty$ , this reduces to minimizing the quantile regression loss function,  $\sum_i \varrho_\tau\{y^{(i)} - (X\boldsymbol{\varphi})^{(i)}\}$ , so the above can be considered a novel form of local polynomial quantile regression. Similar analytical forms of  $g$  obtain also for other forms of  $\psi$ -priors, e. g., Gaussian or uniform ones.



**Fig. 3.** (Colour online) Typical effect of priors for data points placed uniformly at random (black dots) from  $f(x) = \sin(\omega x)$  (thin dashed black line), with  $\omega = 1$  and  $p = 5$ . Left:  $N = 6$ , no value error. Right:  $N = 12$ , medium-sized iid value error with  $\sigma_\varepsilon = 1/4$ . In 100 such samples, the median  $L_2$  distance between true and estimated function in the left[right] case was 0.068[0.24] for proper correlated  $\varphi$ -priors growing as  $\sigma_\varphi^{k_\alpha} \propto \omega^k$  with the correct frequency  $\omega$  (thick solid cyan line), 0.15[0.29] for proper uncorrelated priors of the same size (thick dash-dotted cyan line), 0.22[0.48] for proper correlated priors with  $\sigma_\varphi^{k_\alpha} \propto (\omega/2)^k$  (thin solid blue line), 0.25[1.7] for improper priors (thin dash-dotted red line), and 0.56[0.55] for proper correlated priors with  $\sigma_\varphi^{k_\alpha} \propto (2\omega)^k$  (thin dotted green line).

## 7. Choice of control parameters

### 7.1. Consistent derivatives

With a noninformative  $\varphi$ -prior, the inconsistency between  $\tilde{\mu}_\varphi^{(\alpha)}$  and the  $\alpha$ -th derivative of  $\tilde{\mu}_\varphi^{(0)}$  with respect to  $\xi$  will decrease with  $p$  and vanish for  $p \rightarrow \infty$ . Therefore, if one is interested in all derivatives of order  $\leq q$ , we suggest to use a noninformative prior for all  $\varphi^{(\alpha)}$  with  $|\alpha| \leq q$ , and use a  $p$  somewhat larger than  $q$ , using either informative or noninformative priors for the derivatives of order  $> q$ .

### 7.2. Priors for oscillatory behaviour

*Correlation of derivatives.* Assume that  $d = 1$  and  $f(x)$  is of the form

$$f(x) = \int_0^\infty d\omega F(\omega) \sin(\omega x + t_\omega) \quad (63)$$

with a spectrum  $F(\omega)$  and phases  $t_\omega$  chosen uniformly at random. Then the covariance of  $f^{(k)}(x)$  and  $f^{(\ell)}(x)$  is

$$\Sigma_\varphi^{k,\ell} = \frac{1}{2} \int_0^\infty d\omega F(\omega)^2 \omega^{k+\ell} \cdot \begin{cases} 1 & \text{if } k - \ell = 0 \pmod{4} \\ 0 & \text{if } k - \ell \in \{1, 3\} \pmod{4} \\ -1 & \text{if } k - \ell = 2 \pmod{4} \end{cases} \quad (64)$$

In other words, the  $k$ -th and  $(k + 4)$ -th derivatives are fully correlated, and the  $k$ -th and  $(k + 2)$ -th derivatives are fully anti-correlated (see also Gibson et al. (1992)).

If the spectrum of  $f$  is known approximately, Eq. 64 the above is therefore a natural choice for the  $\varphi$ -prior covariance structure. For unknown spectrum, however, using a prior with strong

correlation between derivatives can result in worse results than using an uncorrelated or improper prior, as can be seen in the example in Fig. 3.

*Growth of derivative variance.* If  $F(\omega)$  is concentrated at a single frequency  $\omega_0$ , i.e.,  $f(x) = F_0 \sin(\omega_0 x + t_0)$ , the variance of  $f^{(k)}(x)$  is growing exponentially:  $\Sigma_\varphi^{k,k} = F_0^2 \omega_0^{2k} / 2$ . For a broader but bounded spectrum, it is growing sub-exponentially. E. g., a uniform spectrum of the form  $F(\omega) = F_0$  for  $\omega \leq \omega_0$  and  $F(\omega) = 0$  for  $\omega > \omega_0$  gives  $\Sigma_\varphi^{k,k} = F_0^2 \omega_0^{2k+1} / (4k+2)$ . For typical unbounded spectra, in contrast, the variance growth is super-exponential. E. g., for an exponentially decaying spectrum with  $F(\omega) = F_0 e^{-a\omega}$  for some  $a > 0$ , as is often assumed of chaotic dynamical systems, we get  $\Sigma_\varphi^{k,k} = F_0^2 (2k-1)!k / (2a)^{2k+1}$ . Similarly, for a Gaussian decaying spectrum with  $F(\omega) = F_0 e^{-\omega^2/2\nu}$  for some  $\nu > 0$ , we get  $\Sigma_\varphi^{k,k} = F_0^2 \nu^{k+1/2} \Gamma(k+1/2) / 4$ . In case of a power-law decaying spectrum, the variance even diverges for large  $k$ . E. g., for  $F(\omega) = F_0(1+a\omega)^{-b}$ , we get  $\Sigma_\varphi^{k,k} = F_0^2 \Gamma(2k+1) \Gamma(2b-2k-1) / 2a^{2k+1} \Gamma(2b)$  for  $k < b-1/2$  and  $\Sigma_\varphi^{k,k} = \infty$  for  $k \geq b-1/2$ , justifying the use of an improper prior.

### 7.3. Data-driven choice of remainder prior

Without argument error, the *posterior predictive distribution* of  $y^{(i)}$  from the Taylor model about  $\xi = x_i$  has

$$\mathbb{E}(y^{(i)}) = \mu_\varphi^{(0)}(\xi = x_i), \quad \text{Var}(y^{(i)}) = \tilde{\Sigma}_\varphi^{(0,0)}(\xi = x_i) + \Sigma_\varepsilon^{(i,i)}. \quad (65)$$

If the model is correct, one therefore expects that approximately

$$\sum_i \frac{|y^{(i)} - \mu_\varphi^{(0)}(\xi = x_i)|^2}{\tilde{\Sigma}_\varphi^{(0,0)}(\xi = x_i) + \Sigma_\varepsilon^{(i,i)}} = N. \quad (66)$$

Likewise, for any fixed  $\xi$ , and if  $P_\varphi = 0$  and  $\mu_\psi = 0$ , the posterior predictive distribution of  $W^{1/2} \mathbf{y}$  from the Taylor model about  $\xi$  has

$$\mathbb{E}(W^{1/2} \mathbf{y}) = W^{1/2} X \tilde{\boldsymbol{\mu}}_\varphi, \quad (67)$$

$$\begin{aligned} \text{Cov}(W^{1/2} \mathbf{y}) &= W^{1/2} X \tilde{P}_\varphi^{-1} X' W^{1/2} + W^{1/2} (\Sigma_r + \Sigma_\varepsilon) W^{1/2} \\ &= W^{1/2} X (X' W X)^{-1} X' W^{1/2} + I. \end{aligned} \quad (68)$$

If the model is correct, one therefore expects that approximately

$$\|W^{1/2} (X \tilde{\boldsymbol{\mu}}_\varphi - \mathbf{y})\|_2^2 = \text{trace}\{W^{1/2} X (X' W X)^{-1} X' W^{1/2} + I\} = m + N. \quad (69)$$

Hence, assuming  $\Sigma_\psi = \nu S$  with a known matrix  $S$  and unknown  $\nu > 0$ , one could either use the same  $\nu$  for all  $\xi$  and choose it so that the “global” Eq. 66 is fulfilled, which is similar to the suggested parameter choice in Wang et al. (2010b), or use a different  $\nu$  for each  $\xi$  and choose it so that the “local” Eq. 69 is fulfilled.

### 7.4. Interpolation with noninformative priors

If  $\Sigma_\varepsilon = 0$  and  $\Sigma_\psi = \nu I$ ,  $\tilde{\boldsymbol{\mu}}_\varphi$  does not depend on  $\nu$ , and Eq. 69 becomes

$$\nu = \|D^{1/2} (X \tilde{\boldsymbol{\mu}}_\varphi - \mathbf{y})\|_2^2 / (m + N), \quad D^{(i,j)} = \delta_{ij} / \sum_{\alpha, |\alpha|=p} (X^{(i,\alpha)})^2 \quad (70)$$

(see Eq. 18). This value is also identical to the posterior mode of  $\nu$  when  $\nu$  is treated as a scale hyperparameter with a noninformative Jeffreys prior of  $\varrho(\nu) \propto 1/\nu$ . Hence the MOTABAR interpolant with noninformative priors is

$$\tilde{\mu}_\varphi = (X'DX)^{-1}X'Dy, \quad \tilde{\Sigma}_\varphi = \frac{\|D^{1/2}(X\tilde{\mu}_\varphi - y)\|_2^2}{m+N}(X'DX)^{-1}. \quad (71)$$

## 8. Example: reconstruction of Lorenz attractor from noisy observations

To demonstrate the performance of MOTABAR for complex data, we simulated a trajectory  $(X, Y, Z)(t)$  of the standard chaotic Lorenz system given by the ODEs  $\dot{X} = 10(Y - X)$ ,  $\dot{Y} = -XZ + 28X - Y$ ,  $\dot{Z} = XY - 10Z/3$ , where  $X$  basically oscillates between  $\approx \pm 15$  with a frequency of  $\approx \omega = 8$  (see Fig. 4,a). We then generated a sample of  $N = 500$  noisy observations  $X(t_i) + \varepsilon(t)$ , with iid errors  $\varepsilon(t) \sim N(0, 1)$  (which corresponds to  $\approx 1\%$  of noise), for time-points  $t_i$  that were regularly spaced at a distance  $\Delta t = 0.05$  (which corresponds to  $\approx 15$  data points per oscillation, see the black dots in Fig. 4,f). From the sample, we reconstructed the original trajectory (Fig. 4,a) up to a diffeomorphism, following the approach of Packard et al. (1980) by estimating the time evolution of the derivatives  $(X, \dot{X}, \ddot{X})$  (Fig. 4,b), using different estimation methods. For MOTABAR estimation, we used  $p = 4$ , an improper  $\varphi$ -prior, and a  $\psi$ -prior whose variance  $\sigma_\psi^2 = 15\omega^p I$  was chosen to fit the approximate range and frequency of the oscillation. Fig. 4(e) shows that the MOTABAR approach reproduces the shape of the trajectory much better than either spline smoothing (c) or numerical differentiation (d), the latter being basically equivalent to an alternative phase space reconstruction method, the often used “method of delays”. Note that there are other, more sophisticated state space reconstruction methods based on PCA or discrete Legendre polynomials, which deal better with noise (Gibson et al., 1992) and which will be compared to MOTABAR in a separate paper.

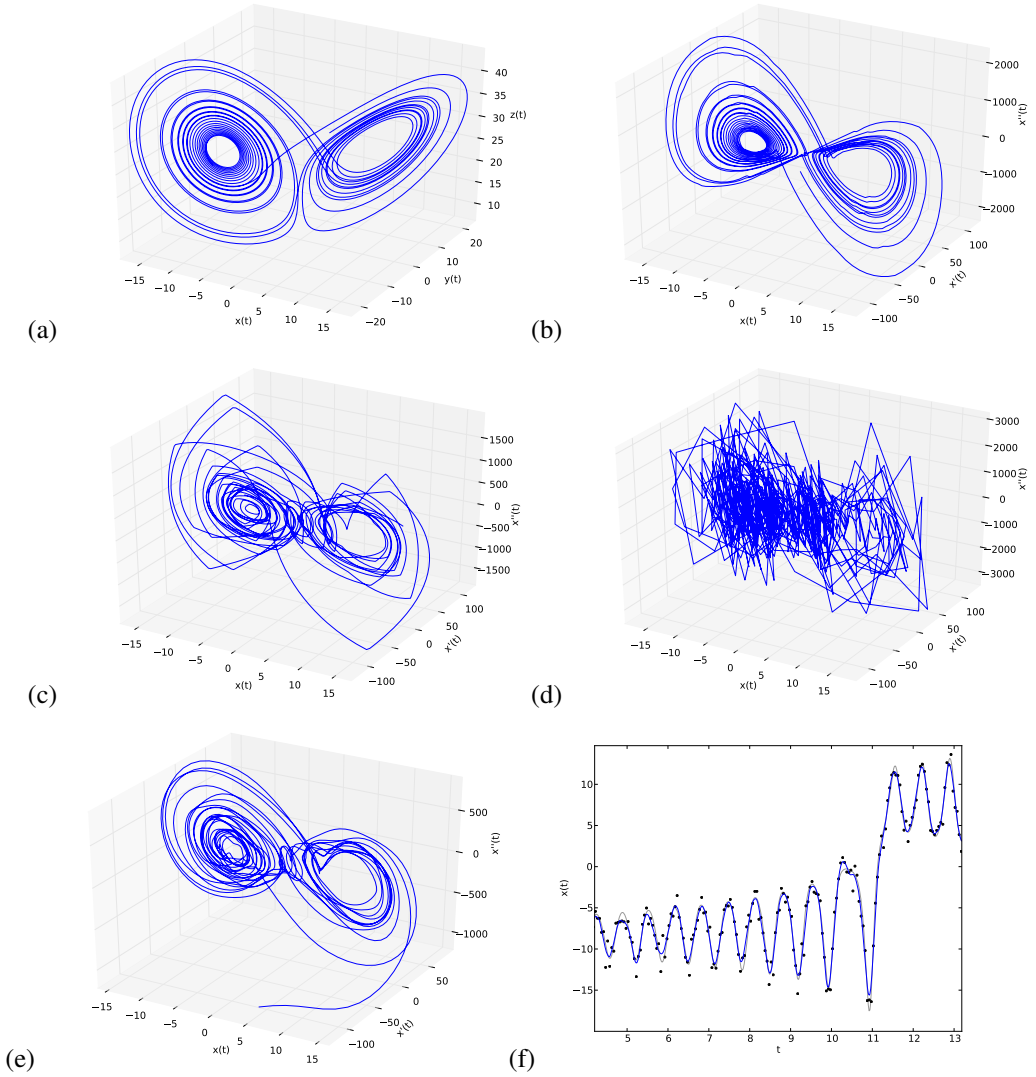
## 9. Conclusion

*Outlook: alternative local approximations.* The simple form of the MOTABAR estimator is due to our assumption of Gaussian value measurement errors and the fact that Taylor polynomials are linear in their coefficients. Alternatively, one might approximate  $f$  by other functions that can be parameterised by the low-order derivatives of  $f$  at  $\xi$ . For  $d = 1$  and if it is known that  $f$  displays oscillatory behaviour, one such approximation could be

$$f(x) = a + b \sin(\omega(x - \xi)) + c \cos(\omega(x - \xi)) + r(x - \xi) \quad (72)$$

with a remainder function  $r$  with  $r(0) = r'(0) = r''(0) = r'''(0) = 0$ . Because then  $\omega = \sqrt{-f'''(\xi)/f'(\xi)}$ ,  $c = f'(\xi)f''(\xi)/f'''(\xi)$ ,  $a = f(\xi) - f'(\xi)f''(\xi)/f'''(\xi)$ , and  $b = f'(\xi)\sqrt{-f'(\xi)/f'''(\xi)}$ , the four derivatives can be estimated from the model if a plausible prior for  $r(x - \xi)$  is used, although the estimate will not be a linear function in  $y$  since the above approximation is not linear in the coefficients. If  $f$  is known to be periodic with frequency  $\omega$ , it might seem that one could also use partial sums of the corresponding Fourier series instead, which are linear in their coefficients, but they are not parameterizable by a finite number of derivatives of  $f$  at  $\xi$ .

*Software.* An open-source software package implementing MOTABAR for use with the python programming language is under development and will be made available at <http://www.pik-potsdam.de/members/heitzig/motabar>.



**Fig. 4.** (Colour online) Reconstruction of the Lorenz attractor from noisy measurements of its  $X$  component. (a) Simulated true phase space trajectory. (b) Derivatives of  $X$  from numerical differentiation from simulated true  $X$  data. (c) Derivatives from natural cubic smoothing spline for noisy measurements of  $X$ . (d) Numerical differentiation from noisy measurements. (e) Derivatives from MOTABAR with  $p = 4$  and an improper  $\varphi$ -prior. (f) Detail of MOTABAR estimated  $X$  (thick blue line) compared to simulated true  $X$  (thin gray line), and noisy measurements (black dots).

## Acknowledgements

This work was supported by the German Federal Ministry for Education and Research (BMBF) via the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS). The author thanks Forest W. Simmons, Kira Rehfeld, Norbert Marwan, Bedartha Goswami, and Jürgen Kurths for fruitful discussions.

## References

- Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829–836.
- Gasca, M. and Sauer, T. (2000) Polynomial interpolation in several variables. *Advances in Computational Mathematics*, **12**, 377–410.
- Gibson, J. F., Farmer, J. D., Casdagli, M. and Eubank, S. (1992) An analytic approach to practical state space reconstruction. *Physica D: Nonlinear Phenomena*, **57**, 1–30.
- Krige, D. G. (1952) A statistical approach to some mine valuation and allied problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **March 1952**, 201–213.
- Kummell, C. H. (1879) Reduction of Observation Equations Which Contain More Than One Observed Quantity. *The Analyst*, **6**, 97.
- Packard, N. H., Crutchfield, J. P., Farmer, J. D. and Shaw R. S. (1980) Geometry from a time series. *Physical Review Letters*, **45**, 712–716.
- Reinsch, C. H. (1967) Smoothing by Spline Functions. *Numerische Mathematik*, **10**, 177–183.
- Shepard, D. (1968) A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference* (eds R. B. Blue and A. M. Rosenberg), pp. 517–524. New York: ACM.
- Sibson, R. (1981) A brief description of natural neighbor interpolation. In *Interpreting Multivariate Data* (ed V. Barnett), pp. 21–36. Chichester: John Wiley.
- Wang, Q., Moin, P. and Iaccarino, G. (2010a) A high order multivariate approximation scheme for scattered data sets. *Journal of Computational Physics*, **229**, 6343–6361.
- Wang, Q., Moin, P. and Iaccarino G. (2010b) A Rational Interpolation Scheme with Superpolynomial Rate of Convergence. *SIAM Journal on Numerical Analysis*, **47**, 4073.